Can Avatars Pass the Turing Test? Intelligent Agent Perception in a 3D Virtual Environment

2 authors:

Richard Gilbert
Loyola Marymount University
27 PUBLICATIONS   **1,019** CITATIONS

SEE PROFILE

Andrew Forney
Loyola Marymount University
12 PUBLICATIONS   **385** CITATIONS

SEE PROFILE

CrossMark

Can avatars pass the Turing test? Intelligent agent perception in a 3D virtual environment ☆

Richard L. Gilbert [a,*], Andrew Forney [b]

[a] The P.R.O.S.E. (Psychological Research on Synthetic Environments Project), Department of Psychology, Loyola Marymount University, USA
[b] Department of Computer Science, University of California, Los Angeles, USA

ARTICLE INFO

ABSTRACT

The current study involved the first natural language, modified Turing Test in a 3D virtual environment. One hundred participants were given an avatar-guided tour of a virtual clothing store housed in the 3D world of Second Life. In half of the cases, a human research assistant controlled the avatar-guide; in the other half, the avatar-guide was a visually indistinguishable virtual agent or "bot" that employed a chat engine called Discourse, a more robust variant of Artificial Intelligence Markup Language (AIML). Both participants and the human research assistant were blind to variations in the controlling agency of the guide. The results indicated that 78% of participants in the artificial intelligence condition incorrectly judged the bot to be human, significantly exceeding the 50% rate that one would expect by chance alone that is used as the criterion for passage of a modified Turing Test. An analysis of participants' decision-making criteria revealed that agency judgments were impacted by both the quality of the AI engine and a number of psychological and contextual factors, including the naivety of participants regarding the possible presence of an intelligent agent, the duration of the trial period, the specificity and structure of the test situation, and the anthropomorphic form and movements of the agent. Thus, passage of the Turing Test is best viewed not as the sole product of advances in artificial intelligence or the operation of psychological and contextual variables, but as a complex process of human–computer interaction.

© 2014 Elsevier Ltd. All rights reserved.

1. The Turing Test: definition and significance

In the now classic paper, "Computing Machinery and Intelligence," Turing (1950) proposed a behavioral approach to determining artificial or machine intelligence. An intelligent agent, or "bot," is said to have passed the Turing Test when it is mistaken by a judge to be a human intelligence in more than 30% of the cases after five minutes of questioning. Procedurally, in what has become known as the "Standard Turning Test" (Sterrett, 2000), a series of judges engage in chat interactions with two agents with the knowledge that one of the agents is human and one is controlled by a computer program. After each judge completes 5 or 10 min of interaction,[1] it is determined whether the percentage of judges who mistakenly identify the computer agent as human exceeds the 30% criterion for classifying the machine as intelligent. In sum, Turing maintained that we will never be able to see inside a machine's hypothetical consciousness, so the best

measure of machine sentience is whether it can fool human judges with some degree of consistency into believing it is human.

Since it was proposed, the Turing Test has remained a source of scientific and public interest and has generated considerable debate regarding both its practical and conceptual significance. Scientists working in the field of artificial intelligence have criticized the practical value of the test by noting that progress in the field does not require its solution (McCarthy, 1996). They maintain that the science of creating intelligent machines (i.e., dynamic machines capable of solving a wide array of problems) no more rests on the ability to create convincing simulations of human mentality than the development of planes rests on how closely they compare to birds (Russell &and Norvig, 2003). At the same time, support for the pragmatic benefits of the test has come from individuals working in the fields of 3D gaming and virtual reality. They argue that the ability to help populate 3D games, virtual reality simulations, and virtual worlds with highly realistic, humanized bots enhances the dynamism of the virtual environment and makes it more compelling for users (Daden Limited, 2010). In addition, the creation of realistic simulations of human mentality brings us closer to the radical possibility of having intelligent agents serve as effective virtual surrogates for human beings and allowing embodied human consciousness, for the first time, to be multiplied in space and time. (Gilbert and Forney, 2013).

---

[1] Some have argued that a test duration of 10 min is needed to comply with Turing's guideline of "after five minutes of questioning," as the judge is only questioning the agents half of the time.

Turing, for his part, was more concerned with the conceptual than the pragmatic implications of his test. His focus was on providing a clear, easily measurable, approach to defining the meaning of artificial intelligence, and thus to provide a specific solution to a demanding philosophical issue. And while many commentators have questioned the validity of his behavioral test as a measure of artificial intelligence, no other proposed definition has proven to be more worthy (French, 1990). In the end, the past half-century has offered no resolution of the pragmatic or conceptual value of the Turing Test and its continued significance may rest more on psychological and existential grounds than on its practicality or conceptual accomplishments. Specifically, the enduring fascination with the Turing Test may be fueled by a deep human interest in understanding the similarities and differences between humans and machines and trying to discern where, if anywhere, the demarcation ultimately lies. Consistent with this view, a number of commentators have maintained that Turing's formulation is less a test of artificial intelligence than it is a modern "indistinguishability test" between humans and machines (Harnad, 1992; Shieber, 2004), a concept first proposed by Descartes in his philosophical treatise, *Discourse on the Method* (Descartes, 1637).

1.1. Efforts to pass the Turing Test

Despite ongoing interest in the Turing Test, no bot has ever successfully passed its original or standard form, although there has been improved performance over time. In recent years, the top performing artificial or computer intelligence programs used in competitions such as the Loebner Prize have been able to deceive judges that they were human in approximately a quarter of the cases (http://www.loebner.net/Prizef/loebner-prize.html). While Turing optimistically predicted that computers with 10 GB of storage would be able to pass the Standard Test by 2000, he seemed to quickly recognize the difficulty of the challenge he had introduced. In a subsequent paper (Turing, 1952), he described a more lenient version of the test, which simply involves a judge asking questions of a computer and deciding whether the computer was controlled by a human or not, that has led to more favorable outcomes.

Using this "Modified Turing Test," Cleverbot, a chat bot created by British scientist, Rollo Carpenter, became the first AI program to pass the Turing Test during the 2011 Techniche festival held in India. Cleverbot is a web-based and mobile application launched in 1997 that uses a growing database of more than 150 million online conversations to chat via text with visitors to its site. Operationally, it finds all keywords that match the input provided by the human with whom it is chatting. It then searches its saved conversations and responds to the input by finding how a human responded to that input, or the most similar input, when it was asked, in part or full, by Cleverbot. In the Techniche competition, Cleverbot was judged to be human by 59.3% of the judges compared to a rating of 63.3% achieved by the human participants (Aron, 2011). In the modified test, a machine needs to reach a deception rate that exceeds 50%, the rate expected by chance alone, rather than the 30% figure that accompanies the Standard Test. This is due to the fact that judges in the modified test only have to make agency decisions based on a single chat interaction as opposed to multiple chat interactions in the standard test.

In addition to moving from the standard to a modified version of the Turing Test, the performance of Cleverbot may have been enhanced by the fact that the judges were naïve to the possibility that they could be interacting with a non-human agent. Various critics have noted that the priming of human interrogators with the possibility that they are interacting with a machine intelligence greatly increases the difficulty of the test (Saygin et al., 2000). They maintain that without this information AI programs would be more successful in deceiving judges that they are human. Consistent with this view, Saygin and Cicekli (2002) employed transcripts from Loebner contestants between 1994 and 1999 to empirically demonstrate that judges are more apt to believe that they are communicating with human beings when they are unaware of the possible presence of machine intelligence.

1.2. Application of the Turing Test to 3D virtual environments

To date, all efforts to pass either the standard or modified version of the Turing Test have taken place in 2D, text-only, environments. The one possible exception occurred during the 2012 Turing centennial when a 3D virtual gamer controlled by an AI program created by computer scientists at The University of Texas won the annual "BotPrize" competition by convincing 52% of judges that it was more human-like than the humans it competed against in a first-person shooter video game (University of Texas, 2012). It should be noted, however, that the BotPrize was awarded based on the ability of the virtual gamer to mimic the behavior of human gamers (e.g., moving around in 3-D space, engaging in chaotic combat against multiple opponents, emulating forms of irrational behavior considered distinctively human) rather than on the type of natural-language processing capabilities emphasized in Turing's formulation of machine intelligence and reflected in the development of Cleverbot. Nevertheless, this finding suggests that 3D virtual environments could be a promising context to attempt to pass a version of the Turing Test, a proposition first advanced by Burden (2008).

Three-dimensional virtual environments have two major advantages as a context to pass the Turing Test relative to text-only programs. First, they are uniquely suited to capitalize on the human tendency toward anthropomorphism, or the process of imbuing something that is non-human with human-like qualities (Epley et al., 2007). Prior research has shown that approximately 95% of avatars (3-dimensional digital representations of the self) created in virtual worlds such as Second Life have a human form (Gilbert et al., 2011). In addition, the digital environment's graphics engine enables avatars to initiate gestures and move through the environment in a human-like manner. It is reasonable to assume that the human form, gestures, and movements of virtual world avatars would strongly activate the anthropomorphic tendency and make it less likely for judges to conclude that an AI program controls them. The second advantage is that interaction between avatars in a 3D environment often takes place in a specific social context (e.g. a club, a store, an art gallery, etc.) where the range of conversational elements are more defined than those in a open-ended, non-contextualized, 2D chat interaction. This contextual restriction serves to reduce the level of natural language processing abilities that are required to persuasively convey a sense of humanness and limits the range of errors that could reveal an avatar's non-human status. Thus, by adding the anthropomorphic and contextual advantages offered by 3D virtual environments to previously identified factors that have increased the likelihood of passing the Turing Test (i.e. using the modified test and judges who are blind to the possible presence of machine intelligence), it is hypothesized that an avatar-mediated Turing Test conducted in a virtual world would result in a highly elevated passage rate.

The current research seeks to test this hypothesis by conducting the first natural language Turing Test in a 3D virtual environment. Specifically, it asks whether individuals operating an avatar in the virtual world of Second Life can accurately determine whether they are interacting with another avatar that is controlled by a human versus machine intelligence if (1) they are naïve to possibility that they could be interacting with an intelligent agent, (2) the interaction occurs in a defined social context, and (3) the human and computer-controlled avatars are visually indistinguishable and equally activate the anthropomorphic assumption. In addition, the present research investigates the temporal aspect

of the Turing Test. While various durations have been used for the test, there is currently no data regarding whether different intervals of interaction significantly impact the accuracy of judgments regarding the presence of human or machine intelligence. As a first examination of this issue, the current research examines the impact of a short and longer form of a modified Turing Test.

2. Methods

2.1. Participants

One hundred participants were recruited via announcements in the Second Life Events Calendar and in high usage regions of the virtual world, notices sent out by heads of large groups representing major constituencies in Second Life, and word-of-mouth communication. Each method of recruitment offered potential participants the opportunity to go on a guided tour of a clothing store located in Second Life and earn 500 Lindens (virtual currency worth approximately $2 US) for completing a brief online survey assessing their impressions of the store and the tour guide. The amount of this incentive is sufficient to purchase a wide variety of items available in the Second Life marketplace and thus, within the virtual context, it reflects a more robust inducement for research participation than a typical incentive of $5–10 for participating in a real-world study. Participants were required to be at least 18 years of age and to have sufficient knowledge of English to participate in an English language tour and survey.

2.2. Procedures

At the start of their scheduled time-slot, participants were teleported by the experimenter to a standardized position in front of the virtual storefront and in close proximity to Ellen, the avatar tour guide. The experimenter then used the local chat function to welcome them, introduce them to Ellen, and assure them that, upon completion of the tour and the ensuing survey, 500 Lindens would be transferred to their accounts. After these preliminaries, the experimenter indicated to Ellen that she could begin the tour of the clothing store. Following this transfer of authority, Ellen instructed participants to follow her to the store and requested that they interact with her in local chat for the remainder of their visit so that a record of their questions and comments about the store could be saved as a chat transcript. In addition, she asked participants to refrain from communicating with other people during the tour (e.g. via instant messaging, email, etc.) in order to maintain their focus on the tour and allow it to be a continuous, uninterrupted experience.

In half of the cases, Ellen was operated by a human controller – a female research assistant who was trained to cover a fixed sequence of topics about the store, ask participants questions about its product line, and answer any questions posed by participants. While conducting the tour, she was instructed to respond naturally to any questions or comments by participants as if she was actually conducting a tour of a clothing store in the physical world. Thus, the research assistant was given a scripted set topics as well as flexibility in responding to participant input.

In the remaining 50 cases, Ellen was a "bot" or intelligent, virtual agent who employed a chat engine called Discourse to interact with participants. As a more robust variant of AIML or Artificial Intelligence Markup Language (Wallace, 2009), the Discourse system allows for rudimentary natural language processing with improved capacity for keyword detection and context specific responses (Burden, 2009).[2]

---

[2] Additional information about the Discourse system can be found at http://www.daden.co.uk/what-is-discourse/.

The bot employs a Second Life "client" script to collect and dispense chat interaction; incoming messages are sent in XML via AJAX calls to a RESTful service operated by Daden, which interprets the input and generates a response. In the current study, various design semantics were added to the bot's scripting to further enhance the human-like feel of the agent, including adding periodic pauses in chat responses, making occasional spelling errors to appear fallible, and shifting to lower-case syntax between or within responses to mimic a human typist (see Appendix C). The bot also maintains a state pertaining to each tour question in order to restrict the domain of expected responses; for example, in the state after asking, "What do you think of the store's color scheme?" the bot is prepared to look for keywords in the participant's answer such as "warm" or "vibrant" in order to respond with more tailored comments such as "Yes, the warm oranges and yellows make me think of summer." Aside from this keyword detection capacity, the bot is otherwise quite restricted, progressing with general answers and sequential script questions regardless of what the participant says. The addition of these humanizing semantics was designed to offset the limitations of AIML in simulating human cognition, including the inability to generalize responses outside of the immediate context of the tour, to respond to clarification questions posed by the participant, or to return to an earlier point in the script if the participant gets distracted, or out of sync, with the interview protocol.

It is important to note that while Ellen was controlled by machine intelligence in the AIML cases, her physical appearance was indistinguishable from the Ellen that was operated by the human research assistant, and participants were not told that the controlling agency of the tour guide was being varied from human to machine across trials. In addition, by having the human and bot-guided tours scheduled on different days or during non-adjacent blocks of time, the human research assistant (as well as the bot, of course) was also unaware that the agency of the tour guide was being varied. Thus, the study involved a double-blind control: both the participants and the research assistant were naïve to variations in the controlling agency of the guide.

The duration of the tour was also varied in order to determine if the accuracy of participant's post-tour judgments of the human or machine agency of the guide was significantly influenced by the length and amount of interaction within the tour. Participants were randomly assigned to receive either a short or long form of the tour, in which the guide posed either ten or fifteen questions over the course of the tour. On average the ten-question tour lasted eight minutes, while the longer tour had an average duration of twelve minutes. Thus, the study consisted of a 2 Agent (human or bot) by 2 Duration (short or long) between-subjects design consisting of four subgroups of 25 participants each.

At the conclusion of the tour, the guide provided participants with a link to an online survey website in order to complete a brief questionnaire assessing their impressions of the store and the tour. The questionnaire largely consisted of a series of 5-point Likert scale questions whose endpoints were 1 (Poor) to 5 (Excellent) that assessed participant's views regarding the esthetics of the store, the presentation of items for sale, and the guide's helpfulness and personal qualities in conducting the tour. All of these questions merely served as precursors to the final two questions that addressed the central issues of the study. Participants were asked to indicate whether they felt that the avatar that guided them through the store was controlled by a human female, a human male, or a computer or artificial intelligence. To eliminate the possibility of order effects, the three response options for the perceived agency of the tour guide were presented in counterbalanced order over the course of the study.

Finally, participants were provided with a text box to explain the factors that led them to determine that the tour guide was under human or computer control. To be specific, participants

were asked to answer the follow-up question: "Briefly, describe what factors made you determine the guide to be either human or computer controlled," after which, the experimenters assessed the presence of three coding criteria of interest: (1) movement, i.e. if the participant spoke of the guide's physical movement as a factor, (2) response, i.e. whether the quality of the guide's responses suggested agency, and (3) the anthropomorphic assumption, i.e. whether the human avatar of the guide begot the initial assumption that its controller was human (see Appendix A).

Following the completion of the survey, participants were thanked for their involvement in the study and 500 Lindens were transferred to their Second Life accounts.

3. Results and discussion

The current study involved the first natural language, modified Turing Test in a 3D virtual environment. As depicted in Table 1, the results indicate that 78% of participants in the artificial intelligence condition (39 of 50) incorrectly judged a bot to be human, while only 10% of participants in the human condition (5 of 50) mistook the human controlled guide for a machine intelligence, reflecting a highly significant difference in the error rate for the bot and human guide conditions, χ^2 (1, $N=100)=46.92$, $p < .001$. Most importantly, the high rate of error found in the bot condition significantly exceeded the 50% rate that one would expect by chance alone that is used as the criterion for passage of a modified Turing Test (Table 2).

A variety of factors contributed to achieving such a high error rate for distinguishing human versus machine intelligence in the bot condition. In the domain of programming, the addition of various humanizing semantics to the bot's scripting (such as adding occasional pauses, spelling errors, shifts to lower-case syntax in chat responses, and generic references to participant comments from previous tours) seemed to aide in the simulation of human cognition and behavior and offset dynamic limitations of the Discourse/AIML engine to generalize responses, clarify questions, or recalibrate following discontinuities in the interview protocol. The positive performance of the chatbot engine was reflected in the responses of participants when

they were asked to explain the basis of their judgments of the tour guide's agency. As depicted in Table 3, 32 of the 39 participants who incorrectly identified the bot as human, or 85%, made some reference to the quality and interactive nature of the bot's responses in shaping their judgments of agency, as reflected in statements such as (the tour guide) "answered questions specifically," "was responsive to my answers," "provided input and appropriate comments," "talked back to me" and "we had a conversation." At the same time, 10 out of 11, or 91%, of the participants who did not attribute human agency to the bot linked their judgment to limitations of the chat engine. Specifically, they cited perceptions that the bot's responses either came too quickly ("she responded too fast," "incredibly fast responses"), seemed overly planned ("the conversation seemed to follow a script," it was "very planned, mechanical"), or it lacked qualities of interaction or responsiveness ("I made a suggestion after the tour and received no response," "I needed more time and she went to another question," "She didn't incorporate or acknowledge what I was saying"). Indeed, the brittleness of the bot's response structure (which simply continues with the interview script lest certain participant keywords appear) compared to its high passage rate further asserts the impact of factors outside of text interaction that contribute to its human-ness. Thus, while the customized chat engine performed credibly, these critical comments highlight some of its dynamic limitations and suggest that factors other than the quality of the chat engine contributed to the high rate in which participants misattributed human agency to the computer-controlled guides.

One of these additional factors is that participants were not given any priming that they could possibly be interacting with an intelligent agent during the tour. As was the case in the Cleverbot demonstration, participants were naïve to the possibility that they could be interacting with a non-human agent, a condition which has been previously found to significantly increase deception rates (Saygin and Cicekli, 2002). The impact of having naïve interlocutors was reflected in participant's responses when they were asked to explain the basis of their judgments of agency. Several participants noted that asking them to explain their decision in assigning agency to the tour guide made them consider the possible presence of a virtual agent for the first time and caused them to reexamine the answers they had given. This was expressed in responses such as "Come to think about it, she may have been computer-controlled," or "Oh…actually, I picked the wrong choice." Thus, it is likely that ensuring that participants were blind to the possible presence of an intelligent agent during the test increased the rate at which bots were misidentified as human-controlled.

It is also likely that the specificity of the context in which the interaction took place increased the rate of human attribution to the bot guides. In contrast to Cleverbot's performance, which occurred in an open-ended, non-contextualized interaction with a relatively high number of conversational exchanges, the current interaction took place in a more defined and restricted context (i.e., a structured tour of a virtual clothing store). The contextual constraints of the test environment lowered the dynamic requirements and natural language processing abilities needed for the bot to appear human and limited the number and type of conversational errors that would signal the operation of machine intelligence. It is interesting to note that among the 11 participants who accurately judged the presence of an intelligent agent, 3 were in the short tour condition (mean duration=8.9 min) and 8 were in the longer tour condition (mean duration=12.3 min). While this does not constitute a significant difference in the rate of successfully identifying the bot as a function of the tour duration (χ^2 (6, $N=100)=8.24$, $p > .05$), it is in the hypothesized direction and may well have risen to the level of significance with a somewhat longer tour or a slightly larger sample. In sum, it is probable that the contextual specificity and temporal restrictions in the current study also played a role in influencing the likelihood that the bot would be incorrectly identified as human.

Table 1
Correctness of agency response as a function of guide controller.

Agency response	Guide controller		χ^2	Φ
	Human	Bot		
Correct	45 (28)	11 (28)	46.92***	.68
Incorrect	5 (22)	39 (22)		

Note: Expected counts appear below frequencies in parentheses.
*** $p < .001$.

Table 2
Correctness of agency response as a function of tour duration.

Agency response	Tour duration		χ^2	Φ
	Short	Long		
Correct	25 (28)	31 (28)	1.461	.31
Incorrect	25 (22)	19 (22)		

Expected counts appear below frequencies in parentheses.

Table 3
Subjective basis of agency responses.

	Agency decision			
	Human trials (50)		Bot trials (50)	
Subjective basis of agency decision[a]	Correct (45)	Incorrect (5)	Correct (11)	Incorrect (39)
Quality of guide responses	37	4	10	32
Quality of guide movement	6	0	2	7
Anthropomorphic assumption	15	2	1	8

[a] *Note*: Some participants indicated more than one basis for their agency decision. Thus the total number of subjective bases exceeds the total number of trials in a category.

The impact of naive interlocutors and contextual constraints are relevant in any platform where an effort is made to pass a modified Turing Test. However, the current study added a unique influence by conducting the test in a 3D virtual environment and using human-form avatars. It was hypothesized that housing the artificial intelligence engine within an avatar whose visual features, gestures, and gross motor movements were human-like would heighten the human tendency toward anthropomorphism and increase the likelihood that the bot would be incorrectly identified as human. Overall, the fact that the avatar-mediated test achieved an unprecedented deception rate strongly suggests that adding anthropomorphic influences to the Turing Test had a significant impact on the current findings. In addition, the important role that anthropomorphism played in the study is reflected in participant's reported reasons for judging the agency of the bot. As shown in Table 3, 15 of the 39 participants who incorrectly identified the bot as human, or 38%, cited the human-form or movements of the avatar, or a pre-existing assumption of human control, as contributing to their judgments of human agency even though anthropomorphic assumptions are generally viewed to operate on an unconscious level (Epley et al., 2007). The explicit operation of anthropomorphic tendencies is reflected in participant statements such as "She seemed human to me in looks, body language, etc.," "The avatar looked female," and "I assumed she was human because I expected her to be" that were used to explain their judgments of agency. Moreover, the fact that 38 of the 39, or 97%, of participants who incorrectly identified the bot as human said it was a Human-Female, the same gender as the tour guide, also indicates that the visual form of the bot was a critical element in determining their judgments. If the visual form of the avatar had played a less central role in participant's decision-making process, one would expect a more equal distribution of gender attributions to the avatar guides that were judged to be human. Thus, the visual features of the avatar guide seemed to exert a strong influence on attributions of gender as well as general human agency.

Taken as a whole, the current findings support a multi-disciplinary perspective on factors that influence the passage of a modified Turing Test, one that combines variables from both computer science and psychology to explain the accuracy of judgments regarding agency. In the domain of computer science, the current findings demonstrate that moderate customization of a basic chat engine such as Discourse/ AIML can contribute to a persuasive sense of humanness in bots equipped with the program when they are situated in a fairly restricted environment. In addition, participants noted several limitations in the bot's responses, such as truncated response latencies and instances of perceived non-responsiveness or interactivity toward participant statements, that could lead to refinements of the bot's algorithms, design semantics, and further increase its realism and believability.

At the same time, the study suggests that the ability of a bot to convey a persuasive sense of humanity is influenced by a set of psychological and contextual variables as well as the sophistication of its artificial intelligence program. Factors such as the naivety of judges to the possible presence of intelligence agents, the duration of the trial period, the specificity and structure of the test situation, and the anthropomorphic form and movements of the agent, all interact with the quality of the AI engine in ultimately determining the human believability of a bot. An artificial intelligence program would require a level of sophistication that surpasses the current state of the art to achieve high deception rates in a modified Turing Test if none of these facilitating psychological and contextual conditions were present. In contrast, as the current study demonstrates, it is possible to achieve deception rates approaching 80% using only a moderately capable chat engine when all of these psychological and contextual factors are favorably represented. Thus, passage of the Turing Test is best viewed, not as the sole product of advances in artificial intelligence or the operation of social-psychological variables, but as a complex process of human–computer interaction.

Future research could help isolate the contribution of artificial intelligence programs, contextual factors, and anthropomorphic tendencies in obtaining high deception rates for Turing Tests conducted within 3D virtual environments. For example, a number of studies could be done comparing deception rates in simple versus complex 3D settings that vary in their context-specific natural language processing demands, or assessing the impact of more elaborate chatbot programs (such as replicating the current design using the Cleverbot engine) that can handle general or tangential conversation. Studies of this nature would increase our understanding of programing and contextual influences on deception rates beyond the impact of the single program and setting used in the current investigation. With respect to anthropomorphic influences, a valuable extension of the current work would be to determine the ability of participants to correctly identify the human or machine agency of the avatar guide if it was presented as an animal, inanimate object, or other non-anthropomorphic form. This would provide additional information regarding the extent to which the human appearance of an avatar influences judgments of agency in a 3D context.

In conclusion, as AI engines increase in sophistication, higher deception rates may be attainable under more demanding social–psychological conditions involving longer trials and more complex settings with greater demands for natural language processing, the current study suggests that 3D virtual environments, a platform that was not even contemplated when Turing first proposed his test, may offer the most favorable context to achieve these challenging outcomes because of their unique ability to activate the anthropomorphic tendency in support of humanizing the computer. Moreover, the facilitating influence of anthropomorphism in 3D environments is likely to intensify as improved 3D graphics engines enable the rendering of avatars with increasingly realistic visual features, gestures, and gross motor movements that strongly suggest a quality of humanness (Dionisio et al., 2013). Looking forward, if advances in the cognitive capacities, visual appearance, and movement skills of 3D bots further blurs the line between human and machine agency, it is likely that robust passage rates for the modified Turing Test will become commonplace in 3D settings and the number of realistic agents and virtual assistants operating in these environments will continue to expand.

Appendix A. Participant subjective response coding examples

Participant response	Assigned condition	Agency correctness	Movement coding	Response coding	Anthro. coding
"Human controlled because of the way the avatar's head moved before she typed. Also responded to my comments with appropriate comments that didn't seem to be scripted."	Human	Correct	Yes	Yes	No
"The avatar looked like a female."	Human	Correct	No	No	Yes
"My own inattention - I assumed she was human because I expected her to be.. Now am unsure"	Bot	Incorrect	No	No	Yes
"The movement of the avatar was not very human. For example, when they led me into the store, they positioned themselves on a very specific location, which is definitely evidence of some sort of navigation script. They also responded way too fast to my input - even for a human using copy/paste!"	Bot	Correct	Yes	Yes	No
"The fact they responded to answers that I gave that wouldn't be able to output by a computer"	Bot	Incorrect	No	Yes	No

Appendix B. Images of the guide (human and bot identical) and tour storefront

See Fig. B1.

Fig. B1.

Appendix C. Sample text interaction transcripts of tour with various guide agency

Human guide

Ellen: Now, if you were walking past this store in Second Life, would you want to visit? Is there anything about the exterior that draws you to it?
Participant A: Basically ;p.
Participant A: Yeah, most likely I would visit.
Participant A: Well, the colors draw my eye.
Participant A: (:
Ellen: What's your opinion of the store's color scheme?
Ellen: oh haha that was my next question

Bot guide

Ellen: Do you think the sparsity of items on this side is easier to browse than the other?
Participant B: The outfits seem sort of similar, except for the far left one
Participant B: Ehhh, the same
Ellen: definitely understandable
Ellen: and are there any other male clothing items that you think might be good to add?
Participant B: Well these are obviously outfits for everyday use....so I'd say perhaps some accessories to go with them would be good. Glasses, hats, prim belts, ect ect.
Ellen: OK we're almost done with the tour. One last question: if you could make one suggestion for improving the store in any way, what might that be?
Participant B: Add more to the male section, have the outfits in here match a bit with the displays outside, have some pictures outside as demos so people can get a general idea of what might be inside of the store.
Ellen: great suggestion, I haven't heard that one before
Ellen: Now, if you were walking past this store in Second Life, would you want to visit? Is there anything about the exterior that draws you to it?
Participant C: the mannequins in the windows draw me in the most, I'd say
Ellen: yes, a lot of people see the mannequins – they must be doing their job well!

Ellen: What's your opinion of the store's color scheme?
Participant C: it works. Earthy colors always go together well.

References

Aron, Jacob, 2011. Software Tricks People Into Thinking it is Human. New Scientist (retrieved 02.10.13).

Burden, D., 2008. Artificial intelligence 2008. In: Paper Presented at the 28th SGAI International Conference on Artificial Intelligence. http://dx.doi.org/10.1016/j.knosys.2008.10.00.

Burden, D., 2009. Emotionally responsive robotic avatars as characters in virtual worlds. In: Proceedings of the 2009 Conference in Games and Virtual Worlds for Serious Applications, 12–19, IEEE Computer Society, Washington, DC, USA.

Daden Limited, 2010. The future of virtual worlds. Unpublished whitepaper.

Descartes, R., 1637. Discourse on Method and Meditations (Laurence J. Lafleur, trans.). The Liberal Arts Press, New York (1960).

Dionisio, J., Burns, W., Gilbert, R., 2013. 3D virtual worlds and the metaverse: current status and future possibilities. ACM Comput. Surv. 45 (3), 1–38.

Epley, N., Waytz, A., Caccioppo, J., 2007. On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. 114 (4), 864–886.

French, R., 1990. Subcognition and the limits of the Turing test. Mind. 99 (393), 53–65.

Gilbert, R., Forney, A., 2013. The distributed self: virtual worlds and the future of human identity. In: Powers, D., Teigland, R. (Eds.), The Immersive Internet: Reflections on the Entangling of the Virtual With Society. Palgrave-Macmillan, pp. 23–37 http://dx.doi.org/10.1057/9781137283023.

Gilbert, R., Foss, J., Murphy, N., 2011. Multiple personality order: physical and personality characteristics of the self, primary avatar, and alt (Springer Series in Immersive Environments). In: Peachey, A., Childs, M. (Eds.), Reinventing Ourselves: Contemporary Concepts of Identity in Online Virtual Worlds, 213–234. Springer Publishing, London http://dx.doi.org/10.1007/978-0-85729-361-9-11.

Harnad, S., 1992. The turing test is not a trick: turing indistinguishability is a scientific criterion. SIGART Bull. 3 (4), 9–10.

McCarthy, J., 1996. From here to human-level AI. In: Paper Presented at the 5th International Conference on Principles of Knowledge Representation and Reasoning, Cambridge, Massachusetts.

Russell, S., Norvig, P., 2003. Artificiai Intelligence: A Modern Approach (2nd. Ed.) Prentice-Hall: Upper Saddle River, New Jersey.

Saygin, A.P., Cicekli, I., 2002. Pragmatics in human–computer conversation. J. Pragmat. 34 (3), 227–258. http://dx.doi.org/10.1016/S0378-2166(02)80001-7.

Saygin, A.P., Cicekli, I., Akman, V., 2000. Turing test: 50 years later. Minds Mach. 10 (4), 463–518. http://dx.doi.org/10.1023/A:1011288000451.

Shieber, S., 2004. The Turing Test: Verbal behavior as a Hallmark of Intelligence. MIT Press, Boston.

Sterrett, S.G., 2000. Turing's two test of intelligence. Minds Mach. 10 (4), 541 http://dx.doi.org/10.1023/A:1011242120015.

Turing, Alan, 1950. Computing machinery and intelligence. Mind LIX 236, 433–460. http://dx.doi.org/10.1093/mind/LIX.236.433.

Turing, Alan, 1952. Can automatic calculating machines be said to think?. In: Copeland, B. Jack (Ed.), The Essential Turing: The Ideas That Gave Birth to the Computer Age. Oxford University Press, Oxford.

Wallace, Richard S., 2009. The Anatomy of A.L.I.C.E.. In: Robert, Epstein, Gary, Roberts, Grace, Beber (Eds.), Parsing the Turing Test. Springer Science+Business Media, London, pp. 181–210.